Award Number: W81XWH-11-2-0226

TITLE: Integrative Lifecourse and Genetic Analysis of Military Working Dogs

PRINCIPAL INVESTIGATOR: Kun Huang, Ph.D.

CONTRACTING ORGANIZATION: The Ohio State University
Columbus, OH 43210-1016

REPORT DATE: October 2013

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
October 2013	Annual	25 October 2012 – 24 October 2013

4. TITLE AND SUBTITLE

Integrative Lifecourse and Genetic Analysis of Military Working Dogs

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-11-2-0226

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Dr. Kun Huang

E-Mail: kun.huang@osumc.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The Ohio State University

Columbus, OH 43210-1016

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The project during the past year accelerated when the CRADA's were executed. In the first two years, we optimized the primary genotyping and molecular methods, and the follow-on validation methods. We also expanded the capabilities of our highly-flexible DAPER database and software tools in the present reporting year. To digitize the pathological records, we initiated the high through-put software customization (ABBYY FlexiCapture) for analysis of 1829 longitudinal veterinary records and AFIP/JPC pathological records. In addition, we have initiated the DoD military dog pathology reports to identify cancer bearing dogs for cancer classification and selection of cases and controls. In the first year we invented an entirely novel approach to conducting genome wide genetic association (GWA) analysis – genomewide IUT analysis (GIA); and in the second year we further validated it. In this second reporting year, we integrated IUT and Bootstrapping as an additional innovation with outstanding utility. Dr. Alvarez's presentation of these methods and results to leaders in the fields of genetics and canine genetics resulted in uniformly positive feedback from them (and multiple requests for collaboration). In addition, Dr. Huang has developed the LDPMap algorithm for enabling accurate query of biomedical terms in the database. In addition, we co-authored (Alvarez) a published study that was not based on the present military dog project, but which made use of the same data mining and analysis methods that will be used in our study. The LDPMap algorithm paper is accepted to BMC Medical Genomics. Dr. Rowell, one of our investigators (originally as a predoctoral student), moved on to conduct a postdoctoral fellowship with a pre-eminent dog geneticist at NIH and, after only a year there, is being recruited for a tenure track faculty position at OSU. Dr. Rybaczyk, another of our investigators (originally a postdoctoral fellow and promoted to research scientist) went on to be an NIH T32 Fellow at MSU, which is essentially a pre-faculty position. Dr. Alvarez was promoted to Associate Professor with tenure by OSU and is now under consideration for leadership training in the OSU College of Medicine.

15. SUBJECT TERMS
none provided

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	58	19b. TELEPHONE NUMBER *(include area code)*

Table of Contents

Page

Introduction

The purpose of this proposal is to provide insight into gene environment interactions. It leverages the simplified genetics and detailed records of the military working dog population. There are several critical aspects to meeting the aims of this proposal. 1) development of data driven selection criteria, 2) biological sampling of representative dogs, and 3) generation of mathematical methodologies capable of handling heterogenous data and statistical tests in consistent manner and providing clear and understandable results that are biologically valid. Here we provide a breakdown of the previous year's work and document our progress towards achieving the specific aims we proposed. While the overall progress of this project is summarized in the Annual Report by Dr. Carlos Alveraz (Lead PI from NCHRI), here are the tasks in which I (Huang from OSU) have engaged in.

Body

Task 1- Regulatory Approval:

i) Cooperative Research And Development Agreements (CRADAs): Both the data and biological CRADAs between Nationwide Children's Hospital (NCHRI; Alvarez, Lead PI, home institution)/OSU (Huang and Couto, Partnering PI's) and DoD/USA were executed by 2013.

ii) Animal use approval (Institutional Animal Care and Use Committee, IACUC): The animal hospital at Lackland AFB received AAALAC accreditation that is mandatory for military IACUC approvals in 2012. In 2013, we submitted final revisions on our IACUC protocol for the collection of biological samples and Lackland veterinary approval was granted; and final Lackland AFB oversight approval was granted and those documents were submitted to DoD CDMRP grant administration. Currently, there is one final approval from ACURO pending (and expected, according to their original anticipated timeline, within ~1 month), at which time biological sample collection can be initiated.

Task 2- Data Capture of Veterinary Records: By having Ms. Michelle Perez, Veterinary Technician, embedded in the military dog health service at Lackland AFB, we have been acquiring clinical and associated data from military dogs. This was made possible by the execution CRADA's in 2013 (Task 1). The veterinary clinical cancer and medical records expertise was provided by Dr. Couto. We have been using that data in two parallel tracks. (i) In the first track, we have been using data forms to create advanced methods for capturing paper-based data and converting those to electronic data (which is classified as raw or manually confirmed to accurately represent the original) (using custom form versions of ABBYY software). That work was initiated in the technical sense before we had CRADA's in place to use it on real DoD military dog health records. In 2013, Mr. Terry Camerlengo and his subsequent replacement Mr. Jacob Aaronson (under supervision of Drs. Alvarez and Huang) worked with actual military dog health records (scanned by Vet. Tech. Ms. Perez at Lackland AFB) to create those custom electronic versions of paper forms. Specifically, they initiated the development of custom scanning and data capture from DoD military dog health record form 1829 (which are generated for each health visit, providing longitudinal data) and from AFIP/JPC pathology reports (which are generated for essentially all diagnostic cancer biopsies and sometimes for necropsy). That required significant efforts from ABBYY support and Research IT, NCHRI to implement. This effort is ongoing. If one or both final customized forms are successful in the near future, we will be able to scan any future records and automatically isolate each 1829 and pathology report. Importantly, we would also be able to scan the many prioritized full records scanned and archived in our database in "track ii". (ii) In the second track that was initiated in 2012 and is ongoing through 2013, we have used different indicators to prioritize individual dogs that are particularly important to our study and have begun scanning their complete records (except for some associated clinical test data that could not be scanned – e.g., EKG's on thin perforated paper (which would have risked their destruction in our portable automatic-feed scanner). We are mainly focused on dogs that have had cancer or most likely would have had it by now if they had high risk (according to age). We thus acquired a list of all Lackland AFB dog health records for which there are AFIP/JPC pathology reports. This was made possible by our primary military dog program contact, LTC Cyle Richard. He provided us that list, which he received from AFIP/JPC; in this way, we did not have to review thousands of records to identify those that contained pathology reports or cancer diagnoses. This in turn allowed us to examine DoD military dog puppy program dog (DoD bred dogs vs. purchased dogs) pedigrees for selection of affected and unaffected littermates or half siblings. From this analysis we identified a relatively small number of popular breeders that had many litters with different partners.

Task 3-Methodolgy Development:

Task 3 is advanced about as far as the data types we have acquired to date. Once final IACUC approval is granted (expected within the month) and we begin to acquire military dog samples after, we expect to be able to deploy the methodologies we have developed. Specifically, we have validated the principal new methods using data from previously-acquired Greyhound osteosarcoma case and control samples, and from data published by the LUPA Consortium (Vaysse A et al. 2011. Identification of genomic regions associated with phenotypic variation between dog breeds using selection mapping. PLoS Genet. 7(10):e1002316. PubMed PMID: 22022279).

In the first year's Annual Report, we included two manuscripts (Rybaczyk et al. and Rowell et al.) that used a new methodology we developed under the present program. Both those manuscripts were submitted for publication in leading genetics journals, and we have been addressing reviewers criticisms and advice. Throughout 2013, we continued to refine and validate those studies. Specifically, this work involves the invention of entirely novel techniques to conduct genomewide association analysis or GWAS (Balding 2006) and multidimensional statistical analysis: Intersection Union

Testing or IUT (Berger 1982; Berger 1997) combined with Bootstrapping (both well established, but the approach has never been used for these applications).

The original focus of these works was on development of the IUT. In the course of improving the methods to address reviewer comments during this reporting year, we determined that the integration of Bootstrapping with IUT is a major innovation and advantage (Fig. 1). The greatest concern about our manuscripts was that the IUT method does not generate conventional measures of statistical significance (p-values), despite the fact that the method empirically ranked IUT-"significant" hits correctly (according to detection of true positives in published datasets). [Notably, that is the major focus of applications of IUT to biology and high throughput gene expression data. Some have proposed solving it using Bayesian approaches, but after many years, no one has had success doing so.] By adding Bootstrapping upstream of IUT, we are able to give another type of measure of robustness of results – a confidence (vs. significance) measure (Bootstrap Confidence Value, BCV).

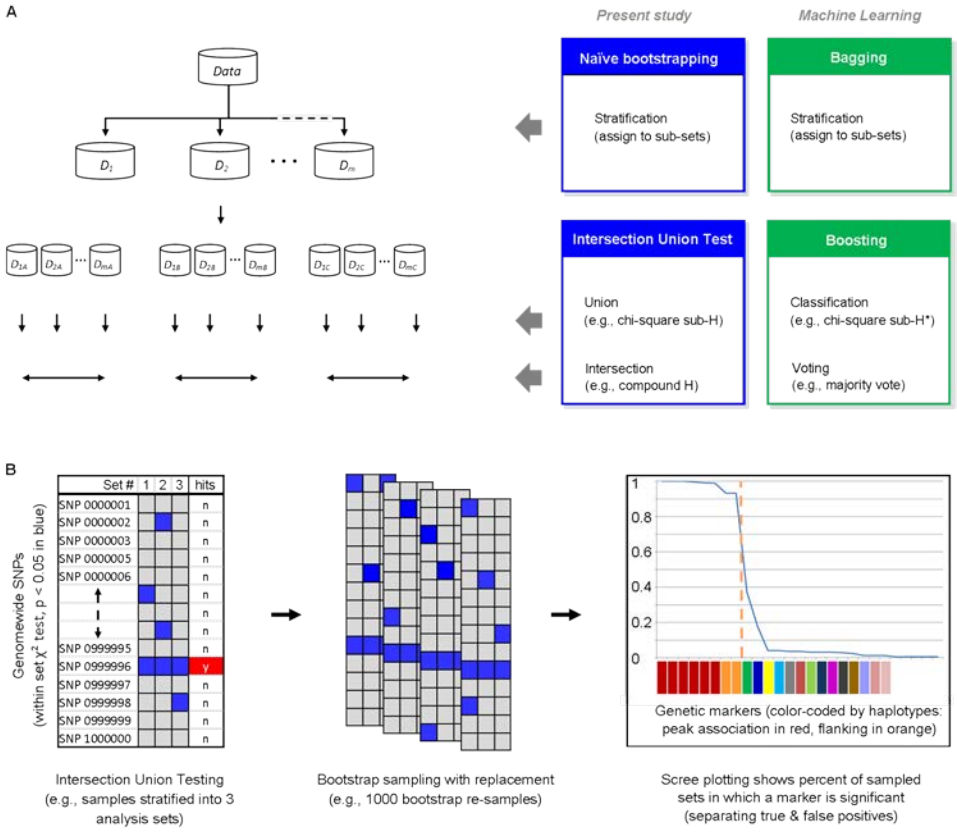

Figure 1. Schematic of integrated Bootstrapping and Intersection Union Testing (IUT) for genetic analysis. (***A***) The schematic on the left shows how a single dataset is repeatedly subsampled (with replacement) and each subsample of cases and controls is then put through the IUT compound hypothesis: i) for each subset of an IUT group, which genetic markers have statistically significant frequency differences in cases and controls, ii) keep only the markers that are significant in all subsets of an IUT (thus not requiring multiple testing correction). Right hand notations compare our methods, which are considered hypothesis tests, to analogous approaches in the field of Machine Learning, which are not considered hypothesis tests but rather learning or predicting. (***B***) Illustration of how IUT works in first panel: each marker (SNP) is tested for significance in each subset of an IUT group (set #1, 2, 3); only those significant in all are kept. Second panel illustrates how repeating the process on 1000 Bootstrap replicates (4 shown) can be used to plot the proportion of times a marker is positive in the 1000 (scree plot, third panel).

In this reporting year we discovered strong evidence that our method is very sensitive and specific based on analysis of the genetic contributions to the complex trait of dog size as a test (using the Vaysse et al. dataset cited above). Specifically, we reanalyzed that published data and, not only identified those authors' two genomewide significant hits using conventional methods, but we also found additional IUT-genomewide significant hits that they missed (but which have been shown to be true positives in other canine genetics studies). We also generated new evidence that Bootstrap/IUT methods i) have increased ability to detect weak signal (a critical need for complex genetics such as cancer risk) and ii) does not require correction for population structure when the analysis is designed properly. We did this by analyzing the most complex dog trait reported by Vaysse et al (ref. above) – sociability (the response of a dog when approached by another dog or a human) as a test (experimental support for these claims were provided in figures within the Q7 and Q8 Quarterly Reports).

In addition to the genetic analysis, we also face the challenge of enabling effective query of medical terms once the database is completed. Given the large collection of biomedical term resources such as ICD9, ICD10, and SNOWMED-CT for clinical diagnosis, Gene Ontology for gene information, and other drug databases, different naming systems can significantly affect the search accuracy. In a collaboration with Dr. Yang Xiang (OSU Biomedical Informatics), we tackle this issue by using the Unified Medical Language System (UMLS) developed by the National Library of Medicine (NLM) of NIH. UMLS has a hierarchical structure for the medical vocabularies collected from more than 100 databases including the ones mentioned above. Each biomedical term is given a unique ID. In order to map the user input words to the exact biomedical terms and IDs in any query, NLM provides a set of tools called Metathesaurus

Browser and MetaMap. However, these tools are quite strict on the input term and often fail if the input term is contains small errors or even small discrepancy with the target term. So we developed a new algorithm called layered dynamic programming mapping (LDPMap) and it provides much higher accuracy in mapping the query terms to the target medical terms. The algorithm was presented in the International Conference on Translational Bioinformatics in Seoul, Korea, in October 2013 and the manuscript was accepted to the special issue for BMC Medical Genomics to be published in 2014 (Ren 2014).

Task 6- Adaptation of existing resources, data storage and hosting:

We have a secure virtual machine called Research DAPER or resdaper developed initially by Mr. Camerlengo and continued by his replacement Mr. Aaronson (supervised by Drs. Alvarez and Huang). The machine exists on the secure NCHRI (Alvarez) network behind a firewall. It can only be accessed by highly-secure VPN using two factor authentication. We have an instance Microsoft SQL Server stored on the machine. Microsoft SQL Server is an industry-leading relational database product that we use to store all of our documents after they have been digitalized. With a relational database, you can quickly compare information because of the arrangement of data in columns. The relational database model takes advantage of this uniformity to build completely new tables out of required information from existing tables. In other words, it uses the relationship of similar data to increase the speed and versatility of the database. The "relational" part of the name comes into play because of mathematical relations. Each table contains a column or columns that other tables can key on to gather information from that table. We have many fields that we can filter and sort on that we can use to retrieve items. Ultimately, this will include all clinical and associated data, environmental data and genetic (genotype), epigenetic, and genomic/molecular (phenotype) data. The user interface is under construction. We will have a web user interface that can be accessed by those with secure credentials. We have used Microsoft asp.net MVC to build the user interface. Using the model view controller pattern gives us the benefit of separating the representation of information from the user's interaction with it .The model consists of application data, business rules, logic, and functions. A view can be any output representation of data, such as a chart or a diagram. The controller mediates input, converting it to commands for the model or view.

In Task 2(i) we discussed the conversion of paper health records to digital versions using ABBYY software – mainly the 1829 form and the AFIP/JPC pathology reports. That digitized data will be fully accessible and searchable through the web interface mentioned above. In addition, the Task 2(ii) scanned complete veterinary clinical records will be directly linked as PDF format. This will allow analysis of digitized data with the option of follow-up detailed analysis of full health records on the same database/tools ensemble “resdaper” (or confirmation/cross-validation of critical data). We have thus installed the ABBYY FlexiCapture software and all of the components which include The Processing Server. That is the server that controls the operation of the Processing Stations. We installed the Licensing Server, the server that stores and manages licenses. We installed the Application Server, the server that controls the operation of the other components. We installed the Application Server components, which will allow operators to connect to the server and work using a web-browser. We also have the Application Server component which allows operators of web stations to register with the system and create requests for access rights to the web station. It provides operators of web stations with a single entry point into the system.

Task 7: Pathway analysis and functional characterization.

Task 7a is complete. I (Alvarez) have been conducting extensive data mining and analysis that are honing those skills which will ultimately be applied to the study of cancer in military dogs. That includes work on osteosarcoma risk candidate genes from Greyhounds (to be published in Rowell et al. manuscript mentioned above) and LUPA candidate genes for multiple canine traits (also discussed above). Most importantly, the Greyhound study implicated small genomic regions with one or two genes each. This allowed use of human cancer data and analysis servers to predict which were likely to be cancer genes and whether the human evidence suggested the cancer risk gene variant was likely to result in up or down regulation. For example, the IntoGen server permits analysis of gene expression and genome alterations associated with diverse cancer types. But other analysis servers, such as NextBio, Oncomine, KMplot and BioGPS provide different tools to mine the same gene expression data in very different ways. For example NextBio make meta-analysis of any subset of studies and KMplot generates Kaplan Meier survival plots for a subset of cancer types that have very large numbers of data available. With this data in hand, it is possible to generate hypotheses and to conduct cross-validation studies. For example, in the Greyhound osteosarcoma case, we can test those predictions by analyzing genetic association candidates in a canine osteosarcoma tumor gene expression dataset which includes Greyhound, Golden Retrievers, Rottweiler’s and mixed breed dogs. Because there are orders of magnitude more human data than canine, it is critical to be able to make use of it.

Among the major aspects of genetic/genomic studies are contextualization according to biochemical or genetic pathways, cross-dimensional/platform validation, and comparative genomics/cross-species validation. To that end, I have

conducted studies in these aspects of cancer genetics. Among those, I mined for genetic evidence that the enzyme aldehyde dehydrogenase is involved in multiple myeloma (for which there is experimental evidence generated by a collaborator studying this with their own funding). As a result of the latter analysis, my analyses were added to a manuscript that was recently accepted for publication. Although the following work was not based on our military dog data, my contributions involve the same analyses that will be conducted with canine cancer candidate genes: Yasmeen R., Meyers J. M., Alvarez C. E., Thomas J. L., Bonnegarde-Bernard A., Alder H., Papenfuss T. L., Benson D. M. Jr, Boyaka P. N., Ziouzenkova O. (2013) Aldehyde dehydrogenase-1a1 induces oncogene suppressor genes in B cell populations. Biochim Biophys Acta 1833:3218–3227. (See Appendix II) For example, I conducted the analysis shown in Figs. 6A and 6C. That critical information shows that the biology suggested by the Yasmeen et al. molecular/biochemical study can be cross-validated by public datasets involving other types of evidence (here gene expression). Similarly, we expect that the vast data available on human cancers will yield supporting evidence for canine cancer findings from the project that is the subject of this report.

Task 8- Project management, Quality control and assurance, and Security:

The most important change in this reporting year is the execution of the CRADA's which allowed us to acquire DoD military dog data. We established a footprint at Lackland and implemented security protocols in accordance with our agreements. We are conducting quality control evaluations for our data collection techniques to assure that we are collecting appropriate data. Once we have assured high quality data we will begin automated import into the database. We are also cross-validating medical and pathology records to assure accurate diagnosis. We initiated collaborations with Dr. David Gutman at Emory University and hope to use his automated pathology data base to facilitate confirmation of sample classification.

As of June 1st, 2013, Task 8 duties attributed to Dr. Rybaczyk (who has moved on in his academic career, as an NIH T32 Fellow, Michigan State U.) are being done by Dr. Alvarez. This transition was been smooth. A job listing was posted for a replacement postdoctoral fellow. Dr. Alvarez interviewed a highly-qualified postdoctoral fellow named Dr. Sohan Lal (currently postdoctoral fellow at Yale), but unfortunately Dr. Lal was forced to accept another position at Yale due to imminent expiration of his visa status. There is another candidate under consideration; the goal is to hire that person prior to initiating the biological sample collection.

The replacement for Mr. Camerlengo – computer programmer – was a success. His role has been taken up by Mr. Jacob Aaronson, who may not be as experienced as Mr. Camerlengo but appears to have greater affinity for the biomedical aspects of computational sciences. In particular, he is a research staff in the Informatics Research & Development Team of the OSU Department of Biomedical Informatics and has extensive experience in developing databases and webtools/interfaces for biomedical applications in the Medical Center. Mr. Aaronson quickly completed his NCHRI orientation, security clearance/ID badge, and vaccination requirements. Most importantly, he rapidly oriented himself in the project and is performing high quality work.

Key Research Accomplishments

- Execution of institutional agreements (CRADA's) between NCHRI (Alvarez)/OSU (Huang, Couto)
- Completion of all facets of IACUC between NCHRI and Lackland AFB through final Lackland AFB oversight approval (currently waiting for final ACURO approval expected within ~1 month)
- Successful embedding of NCHRI (Alvarez) Veterinary Technician, Ms. Michelle Perez within the military dog health service at Lackland AFB
- Successful scanning of veterinary clinical records by Ms. Perez at Lackland AFB, transmission of encrypted data to NCHRI, and uploading to DAPER database
- Continued development and validation of a scale free, high-power statistical methodology capable of resolving signal from noise in high throughput genetic/genomic data (IUT/GIA) by incorporation of Bootstrapping
- GIA manuscripts continue to be refined since receiving comments from peer reviewers
- GIA grant application to NIH is being refined based on peer reviewer critiques
- Expansion of our highly flexible data-infrastructure that is robust enough to handle military working dog records and queries of said records
- Initiation of high through-put software customization (ABBYY FlexiCapture) for analysis of 1829 longitudinal veterinary records and AFIP/JPC pathological records
- Initiation of DoD military dog pathology reports to identify cancer bearing dogs for cancer classification and selection of cases and controls

- Initiation of DoD military dog “puppy program” pedigree analysis for identification of high and low cancer risk lineages
- Development of LDPMap algorithm for mapping query terms to the exact biomedical terms in UMLS.

Reportable Outcomes

- Dr. Jennie Rowell, having received her PhD from OSU for her work at NCHRI (Alvarez), joined the lab of one of two pre-eminent dog geneticists in the world, Elaine Ostrander, NIH, as postdoctoral fellow. The first week of Nov. 2013, she has a job interview for a tenure track position at the College of Nursing, OSU
- Expansion of DAPER database capabilities maintaining strong security
- Mr. Terry Camerlengo moved from OSU to the Battelle Institute as a senior informatics developer. Mr. Jacob Aaronson from OSU Biomedical Informatics IR&D team has successfully replaced Mr. Camerlengo’s role.
- Manuscript for the LDPMap algorithm developed in the collaboration between Dr. Huang and Dr. Xiang is accepted to a special issue in BMC Medical Genomics.

Conclusion

The project accelerated when the CRADA’s were executed. In the first two years, we optimized the primary genotyping and molecular methods, and the follow-on validation methods. We also expanded the capabilities of our highly-flexible DAPER database and software tools in the present reporting year. In the first year we invented an entirely novel approach to conducting genome wide genetic association (GWA) analysis – genomewide IUT analysis (GIA); and in the second year we further validated it. In this second reporting year, we integrated IUT and Bootstrapping as an additional innovation with outstanding utility. Dr. Alvarez’s presentation of these methods and results to leaders in the fields of genetics and canine genetics resulted in uniformly positive feedback from them (and multiple requests for collaboration). In addition, Dr. Huang has developed the LDPMap algorithm for enabling accurate query of biomedical terms in the database. We expect to publish the two revised manuscripts on GIA (one on methods, one on empirical cancer mapping) shortly, but the latter may be delayed while we analyze new supporting data acquired from Dr. Lindblad-Toh. In addition, we co-authored (Alvarez) a published study that was not based on the present military dog project, but which made use of the same data mining and analysis methods that will be used in our study. The LDPMap algorithm paper is accepted to BMC Medical Genomics. Dr. Rowell, one of our investigators (originally as a predoctoral student), moved on to conduct a postdoctoral fellowship with a pre-eminent dog geneticist at NIH and, after only a year there, is being recruited for a tenure track faculty position at OSU. Dr. Rybaczyk, another of our investigators (originally a postdoctoral fellow and promoted to research scientist) went on to be an NIH T32 Fellow at MSU, which is essentially a pre-faculty position. Dr. Alvarez was promoted to Associate Professor with tenure by OSU and is now under consideration for leadership training in the OSU College of Medicine.

References

Balding, D. J. (2006). "A tutorial on statistical methods for population association studies." Nat Rev Genet **7**(10): 781-791.

Berger, R. L. (1982). "Multiparameter Hypothesis Testing and Acceptance Sampling." Technometrics **24**(4): 295-300.

Berger, R. L. (1997). Likelihood Ratio Tests and Intersection-Union Tests. Advances in statistical decision theory and applications. S. Panchapakesan, N. Balakrishnan and S. S. Gupta. Boston, Birkhäuser.

Ren, K., A. Lai, et al. (2014). "Effectively Processing Medical Term Queries on the UMLS Metathesaurus by Layered Dynamic Programming." Accepted to BMC Medical Genomics.

Vaysse, A., A. Ratnakumar, et al. (2011). "Identification of Genomic Regions Associated with Phenotypic Variation between Dog Breeds using Selection Mapping." PLoS Genet **7**(10): e1002316.

Yasmeen R., Meyers J. M., Alvarez C. E., Thomas J. L., Bonnegarde-Bernard A., Alder H., Papenfuss T. L., Benson D. M. Jr, Boyaka P. N., Ziouzenkova O. (2013) Aldehyde dehydrogenase-1a1 induces oncogene suppressor genes in B cell populations. Biochim Biophys Acta 1833:3218–3227.

Appendices

I. Submitted National Institutes of Health grant application including Intersection Union Testing methodology (Aims 2, 3): Statistical techniques for optimized design and power in high-content genomics (Alvarez, PI; Huang, co-PI).

II. Accepted publication: Ren, K., A. Lai, et al. (2014). "Effectively Processing Medical Term Queries on the UMLS Metathesaurus by Layered Dynamic Programming." Accepted to BMC Medical Genomics.

Descriptive Title: Statistical techniques for optimized design and power in high-content genomics

Submission Title:

Opportunity ID: PAR-09-219

Opportunity Title: Exploratory Innovations in Biomedical Computational Science and Technology (R21)

Agency Name: National Institutes of Health

2. SPECIFIC AIMS. This application is in response to PAR-09-219, Exploratory Innovations in Biomedical Computational Science and Technology; it address research, development and application of analytical and statistical tools for interpretation of large biological data sets, and associated software. The flood of biological data has highlighted limitations to signal detection. Here we propose that combining optimized experimental design and novel uses of statistical methods can dramatically increase the power of signal detection. These approaches will be applicable to myriad data types and their integration. However, this proposal will demonstrate validity using a highly innovative approach to complex genetics. We will conduct a Genome Wide Association (GWA) study using high density genotyping that not only provides binary single nucleotide polymorphism (SNP) allele data, but also total SNP signal and allele ratios (which can be affected by DNA copy number variation, CNV). In Preliminary Studies we demonstrate the feasibility of using allele ratios as continuous variables to map disease loci. This is the first such GWA study of comprehensive CNV information without prior classification of markers as CNV. Our ***hypothesis*** is that implementation of our algorithm on multiple (experimentally standardized) groups dramatically increases the power to detect biological signal.

Experimental design. The now common use of thousands or tens of thousands of subjects in genetic studies can be attributed to genetic heterogeneity/complexity and diverse confounds of meta-analysis. A major limitation is the extreme multiple-testing burden in GWA, which is commonly done by Chi-Square testing of one million markers. In Preliminary Studies, we address these issues by 1) conducting complex disease mapping studies in one dog breed, which has 100-fold reduced genetic variation compared to humans, and 2) using multiple, but experimentally identical, case-control sets or batches. In this way, there are reduced numbers of disease-associated markers in a simpler background and we can apply an Intersection Union Test (IUT) across experiments (in place of Bonferroni multiple-test correction). ***Computational statistics.*** The overarching goal of the proposed analytical approaches is based on the information theory concept that the more manipulations or corrections are implemented, the more information is lost. We propose here that this loss of information can be eliminated in diverse types of biological data by integrating two elements. In the first, we use analysis of covariance (ANCOVA) to correct continuous variable data for latent known biological confounders such as group membership. In the second, we make use of optimized study design (specifically, using multiple case-control groups for a given experiment) to perform IUT. Others recently validated a similar use of IUT independently. In Preliminary Studies, we demonstrate validation of the integrated ANCOVA and IUT. We confirm that the use of IUT on multiple sets is a more effective solution to the three reversal paradoxes (Yule-Simpson, Lord's, and suppression) which share the characteristic that the association between two variables can be reversed, diminished, or enhanced when another variable is statistically controlled for. Notably, we are first to address these in the context of continuous genomic variables.

Aim 1: ***Demonstrate on large datasets the ability of ANCOVA to correctly identify biologically relevant phenomena that are linked to a disease trait.*** ANCOVA has been applied to correct for baseline variables in various fields, such as psychology and epidemiology. Despite similarities in variable types, data structure, and confounds, ANCOVA has never been applied to large scale genetic datasets. We will analyze different types of genomic datasets (our own and from the public domain) with well-established population confounds and show that ANCOVA is the most effective way of removing those.

Aim 2: ***Application of IUT for genetic analysis, allowing for multiple corrections without manipulation of individual datasets.*** We propose to demonstrate the ability of IUT to detect complex genetics in a disease phenotype and how combining IUT with ANCOVA will allow the detection of genetic determinants. The non-obvious advancement of this method is that it incorporates information theory by minimally altering the data before analyzing it. This retains the maximum amount of information for each measure. It also does not assume linear relationships with latent variables.

Aim 3: ***We will validate our claim that ANCOVA and IUT are more powerful than traditional techniques.*** We will replicate a published canine complex-genetics mapping study using fewer individuals to demonstrate that our technique is able to detect the same loci in addition other variants missed by traditional techniques. We will also conduct a novel GWA study of a human medically relevant complex trait in a second dog breed.

3. RESEARCH STRATEGY

(a) SIGNIFICANCE

We will develop and implement analytical and statistical tools (and software) for interpretation of large biological data sets. The explosion of biological data has made prominent several limitations to signal detection.[1] We demonstrate in Preliminary Studies that combining optimized experimental design and novel application of statistical approaches can dramatically improve signal detection. These methodologies will be applicable to analytical challenges of myriad data types and their integration [2], including genomics [3], high throughput (HT) sequencing [4], population biology and genetics [5,6], and gene/organism/environment interactions [7]. The improvements described here address the basic concept of information theory that more manipulations of data equals more information loss. Among the areas addressed, are 1) application of analysis of covariance (ANCOVA; [8]) to correct continuous variable data for latent known biological confounders as well as potentially avoiding the three reversal paradoxes (Yule-Simpson, Lord's, and suppression), which share the characteristic that the association between two variables can be reversed, diminished, or enhanced when another variable is statistically controlled for [9,10], and 2) multiple new applications of the Intersection Union Test (IUT; [11]), including GWA, as was independently developed by another investigator very recently [12]. This proposal thus offers solutions and software to address critical barriers to genomic analysis, simultaneously improving scientific knowledge and technical/analytical capabilities.

(b) INNOVATION

Multiple phenotypic traits (such as height or weight) are often treated as independent from the effect under study, but that neglects the reality that many traits are linked to other genetic and environmental modifiers. Others incorporate and calculate variances based on environmental or geographic stratifications. However, this ignores synergism between the organism, its immediate surroundings, and the greater environment. While it is not possible to measure and analyze every part of the environment, some baseline state must be identified from which deviation can be measured to test *a priori* hypotheses. In the absence of this uniform baseline, almost all statistical measures will fail to adequately detect regions of interest. ***This application will demonstrate feasibility and innovation in preliminary studies (c.5) using an entirely new approach (ANCOVA/IUT) to conducting genome wide association (GWA) genetics based on continuous variable data.*** An important challenge to GWA that relates to these issues above is population structure (i.e., correcting genetic studies for non-disease-associated allele frequencies that vary in human populations). Two common ways to address this are traditional meta-analytic techniques and IUT. But these approaches are selected more out of necessity than experimental design concerns. The majority of combinatorial studies have focused on publicly available datasets. Each of the individual datasets contains differing degrees of artifactual bias and other, potentially unrelated, variables. Oncomine's [13] and other algorithms applying this strategy to gene-expression have some success but it has not been the panacea originally prognosticated.[14]

Multivariate and integrative analyses can potentially solve many issues associated with genome wide studies.[15,16] However, they are limited by their ability to synthesize data into useful parcels of information that are applicable clinically or to research. Integrative analysis has the benefit of alternative testing. While multiple testing using the same measures and techniques increases error rates [17], alternative testing allows measurement of the same effect using different types of measures. As these are subjected to different analytic techniques, the posterior probability of false positives is reduced. Even with this strength, it is limited by biases and assumptions associated with individual measures. Ultimately the question of how to appropriately identify genetic contributions independent of latent confounds has not been conclusively answered. The gold standard for analyses is univariate testing. While geneticists talk about penetrance in relation to populations and percentages, the statistical actuality is that penetrance describes odds ratios. Establishing causation and deviation from population norms using case-control, linkage, or association analyses requires certain assumptions to be accepted that biologically may or may not be perilous to the analysis. While this is important to ethologists and population geneticists, attempting to compensate/account for these phenomena hinders and complicates analyses. We are interested in identifying biological outcomes that are well described and were

not concerned with tangential characteristics of the effect. To this end, we sought to isolate rather than compensate for effects. When examining multidimensional data it is easy to disregard the interaction of dimensions. Most dimensional reduction techniques measure and condense data so that interdimensional effects can be quantified. Priming effects can drastically alter these techniques and limit their usefulness. For this reason we applied ANCOVA [8] to remove independent effects from dependent effects prior to dimensional reduction. Here we show adjusted and un-adjusted measures to illustrate how the application of ANCOVA prior to traditional techniques is capable of increasing the sensitivity of a study, as well as the potential to correct for the reversal paradoxes (c.5. P.S., Study Design) by comparison to traditional normalization techniques.

(c) APPROACH

c.1. Research team. The multidisciplinary team is ideally suited for this project. Dr. Alvarez (PI) is PI in Molecular and Human Genetics, Nationwide Children's Hospital Research Institute, with a tenure track academic appointment at The Ohio State University College of Medicine. He has extensive expertise in molecular and human genetics and genomics, bioinformatics, and, from management level industry experience (Novartis Research), the discovery and validation of new drug targets and biomarkers. Dr. Leszek Rybaczyk (Research Scientist, Alvarez Lab) is expert in statistical bioinformatics. Dr. Huang Kun (Co-I) is co-director of the OSU-CCC Biomedical Informatics Shared Resource. His research is focused on developing bioinformatics tools for systems biology and research. Here he will be responsible for developing and implementing the software package. The advanced statistics expertise will come from a long term collaborator of the three investigators named above, Dr. Pramod K. Pathak (consultant, MSU). He is a theoretical and applied statistician with specific interests in statistical methods and their applications to biomedical research, sampling and resampling methods, computational statistics, reliability, and optimization problems in statistics.

c.2. Research strategy (RS). ***Note:*** As the approach has statistical components addressing different biology, we will explain the approach once, in Research Strategy, and establish feasibility in Preliminary Studies.

RS Aim 1. We propose to address these gaps by applying statistically proven methodologies in novel ways. ANCOVA has been applied in various fields such as psychology [18] and epidemiology [19] to correct for baseline variables.[20] Despite the similarities in variable types, data structure, and problems with confounds [19] ANCOVA has never been applied to large scale genetic datasets. **Aim 1:** Demonstrate on a large dataset the ability of ANCOVA to correctly identify biologically relevant phenomena that are linked to a disease trait. The rationale and technical approach for this aim are well elaborated in c.5. Preliminary Studies. Canine genetic data similar to those generated in Preliminary studies will be generated from 1) 36 Scottish Deerhounds: 18 osteosarcoma cases and 18 controls (i.e., three case-control batches of six and six), as well as 2) 36 Doberman (18 with cervical spondylomyelopathy and 18 controls (i.e., three case-control batches of six and six). In addition, we will analyze diverse genomic datasets from the public domain (including human SNP GWA, gene expression, and HT-sequencing). For example, by using TCGA data, in which the same patient's tissue was assayed on different microarrays in different laboratories, using an ANCOVA approach we will identify the most biologically relevant factors. We will expand that by looking not only at the cancer type, but also at the laboratory where the tissue was processed; the date on which it was processed, etc., and identify/potentially remove such intrinsic errors.[21] ***Power analysis.*** Based on our ongoing genetic studies (see Preliminary Studies), we assumed that potentially relevant SNPs will reduce the total of 173,000 SNPs to 1700 [MD Anderson Bioinformatics server with power of 0.8, acceptable false positives of 1, SD of 0.7. With the sample size of 36 dogs in each breed (18 cases and 18 controls) we will have 80 % to detect 2-fold differences in B allele frequency between cases and controls for candidate SNPs of interest (per SNP alpha = 0.00059). This is conservative, as ANCOVA and IUT would only reduce the variance.

RS Aim 1 Potential pitfalls and contingencies. (1) A limitation to using the integrated ANCOVA/IUT on biological data is that it is only applicable for continuous variable data. While this excludes, say, conventional binary-genotype GWA analysis, we address this need with the development of an IUT-alone approach; this use is now validated by us (see c.2. RS Aim 3 Expected results, Example 1) and by a second independent group.[12] Moreover, much genetic data (e.g., array CGH, HT-sequencing) and most genomic data has continuous variables (microarray and HT-sequencing based RNA expression and epigenetics, proteomics, metabolomic, etc.). (2) Another potential concern is the need for clear understanding of appropriate data structure. For that reason, we chose to make this proposal not only about the statistical methods, but also

about experimental design. We will make a major effort to document the proper use of these algorithms in publications and software Help documentation. (3) Lastly, these methods are computationally intensive. This will not affect us, as Dr. Huang (Co-I) is Director of Bioinformatics and has access to the OSU Supercomputer Center. Despite the computational demands, the methods proposed here offer analytical abilities that are unique and state of the art, and are sure to gain wide use. We believe that our optimization studies and careful statistical/software instructions will facilitate the most efficient implementation of our algorithms.

RS Aim 2. A second statistical technique, the Intersection Union Test, has been gaining use in the genomics field.[22] The IUT increases power, but also increases type I error as the number of comparisons increases.[12] However, because of the many latent confounds that cannot be accounted for in most genomic work, the IUT is the most elegant solution to reducing these errors.[23] For instance, in large datasets where a multitude of tests are conducted under traditional techniques, a multi-testing correction would need to be applied. However, as we previously demonstrated using the IUT, the probability of any specific false positive decreases exponentially with the addition of new datasets.[24] This is because the probability of detecting the same false positive in two independent datasets is the multiple of α, traditionally 0.05. For two datasets the probability of the same false positive being detected is 0.0025, for three it is 0.000125, and so on. This can compensate for even large datasets. In datasets with 173,000 variables (SNP arrays used in preliminary studies), using between 4 and 6 independent datasets would eliminate all false positives. Conversely if the same signal is being detected in 6 datasets the probability that it is due to chance is of the order $1.5x10^{-8}$. **Aim 2:** IUT is powerful new tool for genetic analysis and allows for multiple corrections without manipulation of individual datasets. We purpose to demonstrate the ability of IUT to detect complex genetics in a disease phenotype and how combing IUT with ANCOVA will allow the detection of genetic determinants and potentially explain penetrance. The non-obvious advancement of this method is that it incorporates information theory by minimally altering the data before analyzing it. This retains the maximum amount of information for each measure. The IUT is also not hampered by many of the assumptions of other tests.[20]

RS Aim 2 Potential pitfalls and contingencies. The IUT is dependent on having a common variable across all data sets used in the analysis. This variable can be very broad such as dog breed or very narrow such as a molecular phenotype. Regardless, the IUT will only answer questions related to the common variable among data sets. One way to correct for that is in the initial study design. The study design should take into account all of the limitations associated with the various statistical tests a priori. As we recently discussed in a publication, applying the IUT to unrelated data sets will result in the elimination of all signal.[24]

RS Aim 3 rationale. Large scale studies that use traditional GWA require large patient populations to achieve adequate power (and have yet to explain a significant portion of the heritability associated with most diseases).[25,26] This has serious pragmatic and ethical implications.[27] It also poses several experimental design problems as independent irrelevant variables – e.g., in genetics, population structure, can overpower the effect of interest.[28] Manipulation of data by Principal Component Analysis (PCA) after merging, or applying normalizations, hinge on the assumption that the interactions are linear. If the interactions are non-linear, applying these corrections can make analysis more difficult.[29] **Aim 3:** We propose to demonstrate that ANCOVA and IUT are more powerful than the traditional techniques by identifying a study and replicating that study using fewer patients and demonstrating that our technique is able to detect the same signal in addition other variants missed by the more traditional techniques.

RS Aim 3 Genetic studies experimental plan. As we did in Preliminary Studies (c.5., using the same Illumina 173,000 SNP array), we will conduct GWA analysis of two complex traits, each with high incidence in a dog breed. ***Mapping (1)*** As validation of a complex trait that has been mapped using a conventional genetic approach and published, we will map osteosarcoma in Scottish Deerhounds (one locus of dominant effect with evidence of linkage (Z_{max}=5.766)).[30] The original work used a 4-generation pedigree where 60 Deerhounds were genotyped and the genotypes of 70 others were inferred, for a total of 130 dogs. We will replicate that study using the methods developed in this proposal to conduct GWA (ANCOVA/IUT on B allele frequency data and IUT on allele/genotype data) on 18 Deerhound cases and 18 controls (i.e., three case-control batches of six and six). ***Mapping (2)*** In order to immediately draw high impact attention to our innovative approaches, we

propose to conduct GWA of a prominent breed-specific complex-genetic condition with high human relevance – "wobblers" or cervical spondylomyelopathy in Doberman Pinschers (reported to explain 2.5% of proportional mortality in the breed).[31,32] We have been collaborating for over a year with Ronaldo da Costa, our OSU colleague who is a leading authority in this.[32] We are currently conducting pedigree analysis on ~1000 Dobermans (showing strong evidence of heritability; data not shown), and have initiated collection of blood/DNA samples. Using the Doberman wobblers pedigree, we will select optimal informative dogs to conduct a mapping study with 18 cases and 18 controls (i.e., three case-control batches of six and six). ***Power analysis.*** See c.2. RS Aim 1, end of first paragraph.***Follow up to broad mapping:*** depending on the type/strength of the evidence and the length of the haplotypes, we will conduct either fine mapping in related breeds that share a similar phenotype, sequence implicated haplotypes using sequence capture, or characterize transposition events, structural variation or DNA methylation status (see PI (Alvarez) biosketch, which demonstrates successful funding of grants in this area from NIH, DoD CDMRP and AKC-CHF). The PI is expert in genomics and sequence and evolutionary biology analyses that will be required to fully evaluate genetic variants and their possible disease effects.[32-38]

RS Aim 3 Expected results. We predict that in *Mapping (1)* we will identify the same locus published previously (leading to refining the locus through recombination in both breeds), and that we will identify other loci associated with osteosarcoma risk – both SNP alleles and B allele frequency changes suggestive of CNV or of effects resulting in allele-specific SNP genotyping bias from amplification step [[39]]. As Deerhounds are relatively closely related to Greyhounds, we also expect to find some loci shared between the two, which would provide convincing replication of the findings in our preliminary studies. We predict that in *Mapping (2)* we will find wobblers-associated variants. For both mapping studies we expect to identify loci that could not have been found using conventional genetic analyses. *Example 1*, in preliminary GWA studies applying IUT to binary genotype calling of the same Illumina SNP array data used in c.5. Preliminary Studies, we identified a genome wide significant locus that would not have been identified by conventional Chi-Square GWA analysis (not shown). Strikingly, two of the three case-control groups had increased frequency of the SNP allele associated with high risk, but the third group had reduced frequency of the same allele associated with reduced risk. We propose that, due to reversal paradox effects [[9,10]], many such findings cannot be detected by conventional GWA. We also expect to identify candidate genes (e.g., some osteosarcoma candidate haplotypes have no more than one gene) and variants (e.g., through sequence capture) within association loci. *Example 2,* in Preliminary Studies we demonstrate the use of ANCOVA/IUT to identify continuous variable differences in B allele frequencies associated with osteosarcoma risk. This would not be possible with current approaches that map binary SNP alleles (and cannot be detected indirectly by tag-SNPs in LD when the variants are relatively recent). Such variation may be indicative of genetic effects never before sampled genome wide for GWA, such as CNV or isothermal amplification bias [[39]] in Illumina Infinium SNP genotyping (e.g., due to DNA methylation, structural variation, and retrotransposition events). If our expected results materialize, as is strongly supported by our preliminary studies, they would establish the superior power and preservation of information in the innovative experimental design and analyses we propose; and it would open the door to studying the most common (and with highest mutation rates) types of genetic variation [[38]] for the first time.

RS Aim 3 Potential pitfalls and contingencies. Our preliminary studies support the feasibility of applying very well-established statistical methods for novel biological data analyses. For example, applying an IUT approach to GWA using binary genotype data, identified a SNP locus at genome wide significance; but no locus reached significance using conventional Chi-Square analysis on the same genotype data (see Example 1 in previous section). Notably, others have recently independently validated that same application of IUT.[12] A second example is the fact that the ANCOVA/IUT mapping approach identified several loci that were covered by multiple significant SNPs, including five SNPs in a 600,000 kb region of chr6; the odds of the observed physical genome distribution being a random effect are infinitesimally low. The greatest challenges in the field of GWA are validation of association and identification of causative mutations. These remain potential pitfalls for us, but we are encouraged by the fact that our osteosarcoma GWA (using IUT of conventional binary genotypes) in Greyhounds identified one (of 19 significant) SNPs within the 4.5 Mb interval identified for

linkage to osteosarcoma in the closely related Scottish Deerhound. This ability to fine map across related breeds is one of the major strengths of dogs, as are the reduced phenotypic and genetic heterogeneity.[40] For the mutation detection, we will be challenged as is everyone, but 1) we have improved chances over most others because we will have more loci to prioritize for specific molecular approaches based on our types of findings (say, structural variation vs. DNA methylation), and 2) we have the technical and computational expertise, and are using the most cutting edge methodologies.

c.4. Software development

All the algorithms developed in this project will be integrated into an open source R package using R and Bioconductor functions and packages. The package will be tested on both stand-alone workstation and also parallel computing environment including two clusters available at OSU (one in the Ohio Supercomputer Center, one in the Dept. of Biomedical Informatics). The packages will be released on a project website and freely available to public. In addition, we will submit it to Bioconductor in compliance with the testing and inclusion criteria. If time permits, we will also consider integrating the R package into a web tool using web interface tools such as the *Rcgi* package (a CGI WWW interface R).

c.5. Preliminary studies & Demonstration of proposed experimental approach

Note: To demonstrate the novelty and significance, and the experimental plan for all three Aims, we devote significant space in this proposal to describe our preliminary studies (two manuscripts in preparation)..

Study design (ANCOVA/IUT approach), canine osteosarcoma (OSA). Dog breeds have ~100-fold less genetic variation than humans. Greyhounds were split over one hundred years ago into racing and show sub-breeds (registered NGA and AKC, respectively). Strikingly, racers have the highest OSA rate (25% incidence) of any breed, whereas show dogs have no increased risk.[41,42] We thus designed a study of a complex genetic trait in an outbred mammal, but used one of the simplest such contexts possible. Genotyping of these dogs was performed using the highest density SNP array available in dogs (Illumina HD, 173,000 feature; fewer SNPs than humans due to the highly extended linkage disequilibrium (LD) in dogs). Importantly, this genotyping platform provides not only the presence or absence of the binary A or B alleles at each marker, but also the signal intensity of the marker and the ratio of the two alleles (referred to as B allele frequency, BAF). We conducted the SNP genotyping in three OSA positive-negative (case-control) groups in order to 1) using ANCOVA to adjust for group membership as well as potentially addressing the three reversal paradoxes (Yule-Simpson, Lord's, and suppression), which share the characteristic that the association between two variables can be reversed, diminished, or enhanced when another variable is statistically controlled for [[9,10]]; and 2) enable the use of IUT in place of GWA by Chi-Square analysis with Bonferroni multiple testing correction. Specifically, we genotyped batches of 12 dogs in the combination of 4 OSA racers, 4 OSA free racers (OFR) and 4 show (AKC).**Statistics & Results:** Data was analyzed using Illumina GS and Partek GS. Sample attributes (incl.

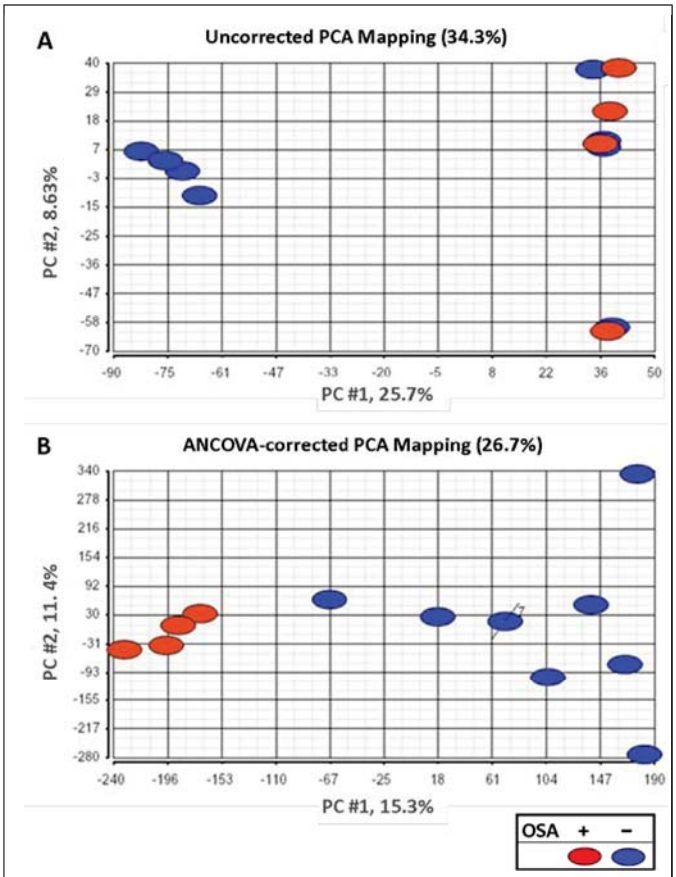

Fig 1. Application of ANCOVA. Correction of Greyhound osteosarcoma (OSA) positive and negative continuous variable genotypes (B allele frequencies). (A) Uncorrected analysis shows population structure effects: separating OSA positive and negative racers apart from negative AKC show Greyhounds. (B) ANCOVA-corrected analysis cleanly separates OSA positive and negative dogs.

racing/show and disease status) were used to assign animals to conditions for ANCOVA corrections. ANCOVA is based on regressions and when used as a statistical test assumes that covariates are independent variables. In our ANCOVA procedure we used it to establish weighted averages so that groups that are biologically similar have the same regression slope. Linear models in biological contexts have been heavily criticized. In this procedure a linear model is entirely appropriate since we are classifying based on known biological traits. Although this does render the measures arbitrary it allows for effects to be isolated that can be subjected post hoc to other tests. Figure 1 demonstrates the effects of ANCOVA isolation on principal components associated with the phenotype of interest. Before correction, two low risk groups (AKC and OFR) fail to cluster according to risk due to population structure. Regression lines were computed for the appropriate factors and interaction values were transformed and weighted to correct for the slope of the generalized linear model. We next calculated the covariance matrix of the loading values for each dataset and conducted IUT using a threshold of ±0.6. Many publications have reported that Pearson correlation (r) values of 0.4 are biologically significant. Here we used 0.6 assuming it most likely captures the most informative SNPs.

Table 1. Analysis of informative SNPs using ANOVA for multiple categories of risk.

SNP	Chr	Position
BICF2S23318678	3	22278940
BICF2P756511	3	34630563
BICF2S22958963	3	34806577
BICF2S23713946	5	3741194
G320f26S259	5	3814438
BICF2P959468	5	24064707
BICF2S23647041	5	25563084
BICF2S23746914	6	71831263
BICF2S22933176	6	72089371
BICF2P643804	6	72282176
BICF2P878053	6	72314083
BICF2S23332924	6	72453644
G439f54S214	7	23851944
TIGRP2P97627	7	49152204
BICF2P989771	9	27058611
BICF2P395540	12	67862864
BICF2P998637	14	39888317
BICF2S23147465	14	51418412
BICF2S2339350	18	23106621
BICF2S23348607	18	23130080
BICF2P950849	18	37553821
TIGRP2P335678	25	54551661
BICF2P691768	28	42235397
BICF2P681391	31	39698895
BICF2P623089	34	30054450

A list of potential candidate SNPs from the ANCOVA/IUT was identified and used to filter genotype information. Genotypes were subjected to a Chi-Square test of association for osteosarcoma risk. Non-significant genotypes were eliminated from the analysis. Once only SNPs that are loaded with the most meaningful measures remained we conducted t-tests to determine if they were capable of discriminating between the two training populations. This procedure revealed that the osteosarcoma free racers and the AKC show greyhounds which have below average incidence rate clustered together and the first principle component explained the osteosarcoma risk variability initially masked by the effects of the population difference (Fig. 1B). We then went on to determine whether it was a genotypic effect such as haplotypes or if some other mechanism was associated with the differential risk in these two populations. Intriguingly, regions associated with altered risk could not be identified based on haplotypes alone. However, the signal was derived from alterations in B allele frequency that correctly categorizing dogs across unrelated datasets. The genome wide significant hits are shown in Table 1. Encouragingly, several regions are detected by multiple SNPs (colored), including five SNPs in a 600,000 kb region of chromosome 6.

Preliminary studies conclusions. Here we presented the first GWA study of osteosarcoma in any organism, and reported approximately twenty hits. Our approach showed how population structure can affect the ability to detect biologically relevant genetic effects. In addition, this is the first work to detect genome wide significant association signal using continuous variable genotype data (B allele ratios) and ANCOVA/IUT; we propose those loci are a combination of CNVs and genetic/epigenetic variants with differing amplification bias [[39]] in the SNP genotyping protocol. This is consistent with Dr. Nadeau's suggestion that the missing heritability may lie in unexplored genome regions or "in largely untested classes of genetic variation."[43] Beyond the analysis shown here, we conducted a second GWA analysis of the same data, but applying only IUT using binary allele calls – see c.2., RS Aim 3, Expected results and Potential pitfalls and contingencies. That analysis suggested validation of the study, as one of 19 genome wide significant hits is within the 4.5 Mb interval linked to osteosarcoma in Deerhounds. Moreover, we identified SNPs that could not be identified by conventional approaches due to the reversal paradoxes.

Application summary: We propose to develop novel applications of validated statistical approaches to enable greatly improved analysis of continuous-variable biological data. This and the new applications of IUT will be widely used for genomic and integrative analyses.

REFERENCES

1. Ideker, T., Dutkowski, J. & Hood, L. Boosting signal-to-noise in complex biology: prior knowledge is power. *Cell* **144**, 860-3 (2011).
2. Hawkins, R.D., Hon, G.C. & Ren, B. Next-generation genomics: an integrative approach. *Nat Rev Genet* **11**, 476-86 (2010).
3. Kim, K., Zakharkin, S.O. & Allison, D.B. Expectations, validity, and reality in gene expression profiling. *J Clin Epidemiol* **63**, 950-9 (2010).
4. Schadt, E.E., Linderman, M.D., Sorenson, J., Lee, L. & Nolan, G.P. Computational solutions to large-scale data management and analysis. *Nat Rev Genet* **11**, 647-57 (2010).
5. Sato, Y., Laird, N.M. & Yoshida, T. Biostatistic tools in pharmacogenomics--advances, challenges, potential. *Curr Pharm Des* **16**, 2232-40 (2010).
6. McCarthy, M.I. et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* **9**, 356-69 (2008).
7. Lewontin, R. *The Triple Helix: gene, organism and environment*, (Harvard University Press, Cambridge, MA, 2000).
8. Huitema, B.E. *The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies*, (Wiley, 2011).
9. Blyth, C.R. On Simpson's Paradox and the Sure-Thing Principle. *Journal of the American Statistical Association* **67**, 364-366 (1972).
10. Tu, Y.K., Gunnell, D. & Gilthorpe, M.S. Simpson's Paradox, Lord's Paradox, and Suppression Effects are the same phenomenon--the reversal paradox. *Emerg Themes Epidemiol* **5**, 2 (2008).
11. Berger, R.L. Multiparameter hypothesis testing and acceptance sampling. *Technometrics* **24**, 295-300 (1982).
12. Shriner, D. & Vaughan, L.K. A unified framework for multi-locus association analysis of both common and rare variants. *BMC Genomics* **12**, 89 (2011).
13. Rhodes, D.R. et al. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia* **9**, 166-80 (2007).
14. Gadbury, G.L., Garrett, K.A. & Allison, D.B. Challenges and approaches to statistical design and inference in high-dimensional investigations. *Methods Mol Biol* **553**, 181-206 (2009).
15. Xie, Y. & Ahn, C. Statistical methods for integrating multiple types of high-throughput data. *Methods Mol Biol* **620**, 511-29 (2010).
16. Peng, J. et al. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Ann. Appl. Stat.* **4**, 53-77 (2010).
17. Sainani, K.L. The problem of multiple testing. *PM R* **1**, 1098-103 (2009).
18. Adams, K.M., Brown, G.G. & Grant, I. Analysis of covariance as a remedy for demographic mismatch of research subject groups: some sobering simulations. *J Clin Exp Neuropsychol* **7**, 445-62 (1985).
19. Koch, G.G., Tangen, C.M., Jung, J.W. & Amara, I.A. Issues for covariance analysis of dichotomous and ordered categorical data from randomized clinical trials and non-parametric strategies for addressing them. *Stat Med* **17**, 1863-92 (1998).
20. Little, R.J., An, H., Johanns, J. & Giordani, B. A comparison of subset selection and analysis of covariance for the adjustment of confounders. *Psychol Methods* **5**, 459-76 (2000).
21. Leek, J.T. et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* **11**, 733-9 (2010).
22. Thorrez, L. et al. Tissue-specific disallowance of housekeeping genes: the other face of cell differentiation. *Genome Res* **21**, 95-105 (2011).
23. Westfall, P.H., Ho, S.Y. & Prillaman, B.A. Properties of multiple intersection-union tests for multiple endpoints in combination therapy trials. *J Biopharm Stat* **11**, 125-38 (2001).
24. Rybaczyk, L.A., Bashaw, M.J., Pathak, D.R. & Huang, K. An indicator of cancer: downregulation of monoamine oxidase-A in multiple organs and species. *BMC Genomics* **9**, 134 (2008).
25. Pasaniuc, B. et al. Enhanced statistical tests for GWAS in admixed populations: assessment using African Americans from CARe and a Breast Cancer Consortium. *PLoS Genet* **7**, e1001371 (2011).
26. Manolio, T.A. et al. Finding the missing heritability of complex diseases. *Nature* **461**, 747-53 (2009).
27. Kaye, J., Boddington, P., de Vries, J., Hawkins, N. & Melham, K. Ethical implications of the use of whole genome methods in medical research. *Eur J Hum Genet* **18**, 398-403 (2010).

28. Cardon, L.R. & Palmer, L.J. Population stratification and spurious allelic association. *The Lancet* **361**, 598-604 (2003).
29. Rybaczyk, L.A. Ohio State University (2008).
30. Phillips, J.C., Lembcke, L. & Chamberlin, T. A novel locus for canine osteosarcoma (OSA1) maps to CFA34, the canine orthologue of human 3q26. *Genomics* **96**, 220-7 (2010).
31. Mandigers, P.J., Senders, T. & Rothuizen, J. Morbidity and mortality in 928 Dobermanns born in the Netherlands between 1993 and 1999. *Vet Rec* **158**, 226-9 (2006).
32. da Costa, R.C. Cervical spondylomyelopathy (wobbler syndrome) in dogs. *Vet Clin North Am Small Anim Pract* **40**, 881-913 (2010).
33. Alvarez, C.E. On the origins of arrestin and rhodopsin. *BMC Evol Biol* **8**, 222 (2008).
34. Alvarez, C.E., Robison, K. & Gilbert, W. Novel Gq alpha isoform is a candidate transducer of rhodopsin signaling in a Drosophila testes-autonomous pacemaker. *Proc Natl Acad Sci U S A* **93**, 12278-82 (1996).
35. Alvarez, C.E., Sutcliffe, J.G. & Thomas, E.A. Novel isoform of insulin receptor substrate p53/p58 is generated by alternative splicing in the CRIB/SH3-binding region. *J Biol Chem* **277**, 24728-34 (2002).
36. Chari, R. et al. SIGMA: a system for integrative genomic microarray analysis of cancer genomes. *BMC Genomics* **7**, 324 (2006).
37. Chen, W.K., Swartz, J.D., Rush, L.J. & Alvarez, C.E. Mapping DNA structural variation in dogs. *Genome Res* **19**, 500-9 (2009).
38. Alvarez, C.E., Akey, J. M. Copy Number Variation in the Domestic Dog. *Mamm Genome* **In press**(2011).
39. Schaerli, Y. et al. Isothermal DNA amplification using the T4 replisome: circular nicking endonuclease-dependent amplification and primase-based whole-genome amplification. *Nucleic Acids Res* **38**, e201 (2010).
40. Rowell, J.L., McCarthy, D.O. & Alvarez, C.E. Dog models of naturally occurring cancer. *Trends Mol Med* (2011).
41. Lord, L.K., Yaissle, J.E., Marin, L. & Couto, C.G. Results of a web-based health survey of retired racing Greyhounds. *J Vet Intern Med* **21**, 1243-50 (2007).
42. Zaldivar, S., Marin, L.M., Hamilton, H. & Couto, C.G. Research Abstract Program of the 2009 ACVIM Forum & Canadian Veterinary Medical Association Convention. *Journal of Veterinary Internal Medicine* **23**, 673-786 (2009).
43. Eichler, E.E. et al. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* **11**, 446-50 (2010).

Effectively processing medical term queries on the UMLS Metathesaurus by layered dynamic programming

Kaiyu Ren[1,2], Albert M. Lai[1], Aveek Mukhopadhyay[2], Raghu Machiraju[2], Kun Huang[1], Yang Xiang[1§]

[1] Department of Biomedical Informatics, the Ohio State University, Columbus, OH 43210, USA

[2] Department of Computer Science and Engineer, the Ohio State University, Columbus, OH 43210, USA

Contact: Yang Xiang

[§]Corresponding author

Email addresses:

YX: yxiang@bmi.osu.edu

Abstract

Background

Mapping medical terms to standardized UMLS concepts is a basic step for leveraging biomedical texts in data management and analysis. However, available methods and tools have major limitations in handling queries over the UMLS Metathesaurus that contain inaccurate query terms, which frequently appear in real world applications.

Methods

To provide a practical solution for this task, we propose a layered dynamic programming mapping (LDPMap) approach, which can efficiently handle these queries. LDPMap uses indexing and two layers of dynamic programming techniques to efficiently map a biomedical term to a UMLS concept.

Results

Our empirical study shows that LDPMap achieves much faster query speeds than LCS. In comparison to the UMLS Metathesaurus Browser and MetaMap, LDPMap is much more effective in querying the UMLS Metathesaurus for inaccurately spelled medical terms, long medical terms, and medical terms with special characters.

Conclusions

These results demonstrate that LDPMap is an efficient and effective method for mapping medical terms to the UMLS Metathesaurus.

Background

Efficiently processing and managing biomedical text data is one of the major tasks in many medical informatics applications. Biomedical text analysis tools, such as MetaMap [1] and cTAKES [2], have been developed to extract and analyze medical terms from biomedical text. However, medical terms often have multiple names, which make the analysis difficult. As an effort to standardize medical terms, the

Unified Medical Language Systems (UMLS) [3] maintains a very valuable resource of controlled vocabularies. It contains over 200 million medical terms (also known as "medical concepts"). Each medical term is identified by a unique id known as a Concept Unique Identifier (CUI). The UMLS also records relations between medical terms. As a result, mapping biomedical text data to the UMLS and mining UMLS associated datasets often yield rich knowledge for many biomedical applications [4] [5] [6] [7] [8].

In order to effectively query or use the UMLS, one of the fundamental tasks is to correctly map a biomedical term to a UMLS concept. Currently, there are a number of publicly available tools to achieve this goal. One notable approach is to use the official UMLS UTS service (UMLS Metathesaurus Browser) available on the UMLS official website (https://uts.nlm.nih.gov). Users are able to input a medical term and the system will return a query result. MetaMap [1], which has been developed and maintained by US National Library of Medicine, has become a standard tool in mapping biomedical text to the UMLS Metathesaurus. cTAKES [2] is an open-source natural language processing system that can process clinical notes and identify named entities from various dictionaries, including the UMLS.

However, after having been using these tools in our research, we found that they do not work well in mapping medical terms that are just slightly different from the terms in the UMLS. For example, the UMLS Metathesaurus Browser, MetaMap, and cTAKES fail to process the query term "1-undecene-1-O-beta 2',3',4',6'-tetraacetyl glucopyranoside" even if it has only one character different (missing "-" between "beta" and "2") from the official UMLS concept "1-undecene-1-O-beta-2',3',4',6'-tetraacetyl glucopyranoside". This drawback makes it hard to handle many real world data such as Electronic Health Records, which contain a lot of noisy information

including missing and incorrect data [9]. In addition, they often fail to handle long medical terms even if those terms are identical to the terms in the UMLS. For example, the Metathesaurus Browser cannot handle query terms with more than 75 characters, and sometimes cannot even accurately answer a query term that exactly matches a concept name in the UMLS (see discussions in the result section). MetaMap and cTAKES, on the other hand, often breaks down a long medical term into several shorter terms. For example, if we query MetaMap with a clinical drug "POMEGRANATE FRUIT EXTRACT 150 MG Oral Capsule", we get several UMLS concepts such as "C1509685 POMEGRANATE FRUIT EXTRACT", "C2346927 Mg++", and "C0442027 Oral", instead of this drug concept which has a unique CUI C3267394 in the UMLS. The situation becomes even worse when medical terms contain special characters, i.e., characters other than numbers or letters, such as "{", "}", " (",")","-", etc. For example, MetaMap completely fails to find any relevant CUI to the medical concept "cyclo(Glu(OBz)-Sar-Gly-(N-cyclohexyl)Gly)2". These drawbacks are very undesirable when handling biomedical texts. By studying the UMLS Metathesaurus, we found that a significant number of medical terms are quite long. About 10.7% of UMLS concepts contain at least 75 characters (including white spaces), and about 50.9% of UMLS concepts contains at least 32 characters. In addition, a large amount of medical terms contain special characters. More than 61.3% of UMLS concepts contain at least one special characters and about 11% of UMLS concepts contains at least 5 special characters. In fact, we found many special characters are optional in a medical term. For example, term “Cyclic AMP-Responsive DNA-Binding Protein” and term “Cyclic AMP Responsive DNA Binding Protein” both refer to the same concept "C0056695" in the UMLS Metathesaurus, though the latter is missing two "-". The UMLS handles a medical term with different

names by including multiple common names in the Metathesaurus. Given the fact that in many cases special characters are optional, it is practically impossible to let Metathesaurus contain all possible names. Considering a UMLS concept with 20 special characters, if each special character may be replaced by a white space, then there are approximately 1 million aliases for this concept alone, not to mention that more than 0.3% of UMLS concepts contain 20 special characters or more.

This problem is in fact related to the classical spelling correction problem in which a misspelled word will be corrected to the most closely matched word. The classic measurement of dissimilarity between two words based on several distance functions, such as edit distance [10], hamming distance [11], and longest common subsequence distance [12] [13]. Thus the spelling correction is essentially finding a valid word with the minimum distance to the misspelled word. Quite a few dynamic programming algorithms have been proposed to solve this problem. Readers can find a survey of these algorithms in [14]. In recent years, spelling correction has evolved to perform query corrections. This correction is often a task of context sensitive spelling correction (CSSC), where corrections will be geared towards more meaningful or frequently searched words [15]. Thus, it is a good idea to use the query log to assist the correction [16].

Unlike many query applications, it is not sufficient to return a frequently searched medical term that best matches the query based on search history, not to mention that such history data is often not available. Accurately identifying a specific biomedical term, such as a drug name or a chemical compound, is demanded by many biomedical applications. Given this consideration, classical spelling correction techniques are more preferable than the CSSC for matching biomedical terms to UMLS concepts. However, we found that the classical dynamic programming algorithm is too slow for

this task because of the huge volume of terms in the UMLS Metathesaurus. In addition, it is unable to effectively handle a term with missing words (e.g., "gastro reflux" has a large distance to "gastro oesophageal reflux" though the two terms usually means the same thing), or words not in their usual order (e.g., "lymphocytic leukemia chronic" has a large distance to "leukemia chronic lymphocytic").

The background described above motivated us to find an efficient and accurate medical term mapping method for the UMLS. To tackle this challenge, in this work we propose a Layered Dynamic Programming Mapping (LDPMap) approach to query the UMLS Metathesaurus.

Methods

We use Longest Common Subsequence (LCS) to measure the similarity between two words. Given two words *A* and *B*, their similarity is defined as:

$$WordSimilarity(A, B) = 2*|LCS(A,B)| / (|A|+|B|);$$

This similarity measure is a variation of the longest common subsequence distance [12]. We can observe that $WordSimilarity(A, B)$ ranges between 0 and 1. In addition, $WordSimilarity(A, B) = 1$ if and only if *A* and *B* are identical, and $WordSimilarity(A, B) = 0$ if and only if *A* and *B* shares no common letters.

The function $WordSimilarity(A, B)$ is the basic building block for LDPMap. In the UMLS, each concept is a sequence of words. We define the similarity between two concepts $\alpha_n = (A_1, A_2, \ldots, A_n)$ and $\beta_m = (B_1, B_2, \ldots, B_m)$ as:

$$ConceptSimilarity(\alpha_n, \beta_m) = max\left(\sum_{(i,j)\in R} WordSimilarity(A_i, B_j)\right);$$

Similar to word similarity, in our query we will normalize the concept similarity by the number of words contained in each concept. We can observe that normalized

concept similarity score ranges between 0 and 1. If two concepts are identical then this score is 1.

$$NormConceptSimilarity(\alpha_n, \beta_m) = 2*ConceptSimilarity(\alpha_n, \beta_m)/((|\alpha_n| + |\beta_m|);$$

The key issue in the above definition is *R*, which is a matching relation between words in α and β. We have two constraints on *R*, which leads to two different foci.

Constraint 1: There do not exists two matching pairs (*i*,*j*), (*x*,*y*) in *R* such that $i=x$ or $j=y$.

Constraint 2: In addition to constraint 1, for any two matching pairs (*i*,*j*), (*x*,*y*) in *R*, either $i<x$ && $j<y$, or $x<i$ && $y<j$.

Constraint 1 converts the concept similarity problem into a maximum weighted bipartite matching problem [17]. Considering a bipartite graph built on two vertex sets α_n and β_m with word similarities being the edge weights, finding a highest score for concept similarity under Constraint 1 is equivalent to find a maximum weighted matching for the bipartite graph. This model is particularly helpful for identifying the similarity between two terms regardless of their word ordering. We used this as one of the measurements in our final query workflow (Figure 1) and implemented this by maximal weighted matching.

In the following section, we will focus on concept similarity calculation under constraint 2, which regulates that the similarity comparison between two terms shall follow the word orders in those terms, similar to the LCS problem in which matching between two words shall follow the character orders. Thus, the concept similarity calculation problem can be considered as a macro level similarity calculation where each unit is a word instead of a letter as in the case of word similarity calculation. This model has a lot of advantages as we will see in the following section.

Suboptimal Structure of the Concept Similarity under Constraint 2

Our next question is how to perform the concept similarity calculation. Unlike word similarity calculation in which each match outcome is a binary result (i.e., the same letter or a different letter), each match in the concept similarity calculation is a word similarity value between 0 and 1. The algorithm for the word similarity calculation cannot be applied to the concept similarity calculation. However, we find the concept similarity calculation also has a suboptimal structure as follows:

if i=0 or j=0

$ConceptSimilarity(\alpha_i, \beta_j) = 0$

else

$ConceptSimilarity(\alpha_i, \beta_j) = max(ConceptSimilarity(\alpha_{i-1}, \beta_{j-1}) + WordSimilarity(A_i, B_j), ConceptSimilarity(\alpha_i, \beta_{j-1}), ConceptSimilarity(\alpha_{i-1}, \beta_j))$;

The above suboptimal structure is true because for any two words $A_i \in \alpha_i$, $B_j \in \beta_j$, there are at most three possible cases:

(1) $(i, j) \in R$, i.e, Both A_i and B_j are used in the matching. Then $ConceptSimilarity(\alpha_i, \beta_j) = ConceptSimilarity(\alpha_{i-1}, \beta_{j-1}) + WordSimilarity(A_i, B_j)$;

(2) B_j is not used in the matching, then $ConceptSimilarity(\alpha_i, \beta_j) = ConceptSimilarity(\alpha_i, \beta_{j-1})$;

(3) A_i is not used in the matching, then $ConceptSimilarity(\alpha_i, \beta_j) = ConceptSimilarity(\alpha_{i-1}, \beta_j)$.

Note that we do not consider it a valid case that neither A_i nor B_j is used in the matching. In this case, we can always choose to make them matching without violating Constraint 1 and result in a higher or at least equal concept similarity score.

Main Algorithms

Given the suboptimal substructure, we can design a dynamic programming algorithm to calculate the concept similarity score between two terms, on top of the LCS dynamic programming algorithm for calculating word similarity. The two layers of dynamic programming not only result in a method less affected by missing words or words in different orders, but also significantly increase the query speed as we will see below. These enable our searching method practically applicable to many biomedical applications.

The UMLS Metathesaurus (version used in this work: 2012AB) contains around 11 million records in its MRCONSO.RRF files. Each record is a medical term. For query purposes, we discard duplicate terms and non-English terms and result in about 6.87 million records. A term is considered duplicate if both its CUI and name are identical to another term. However, among these 6.87 million records, there are only 1,874,573 unique words (white space is the delimiter). Thus concept similarity on a word basis saves a huge amount of redundant calculation otherwise needed by classic methods on a character basis. Correspondingly, in our method, we first pre-process the UMLS Metathesaurus into a word vector of unique words, and convert each UMLS concept, which consists of a list of words, into a list of indices with regard to the word vector. Procedure LDPMap-Preprocessing is the pseudo code.

Procedure LDPMap-Preprocessing ( )

1: **for** *i*=1: length (*Metathesaurus*)

2: *Word_Vector* = *Word_Vector* ∪ *Metathesaurus* [*i*];

3: **endfor**

4: **for** *i*=1: length (*Metathesaurus*)

5: **for** *j*=1: length (*Metathesaurus*[*i*])

WordIndex_vector [*i, j*] = the index of *Metathesaurus*[*i, j*] in *Word_Vector*;

6: **endfor**

7: **endfor**

8: **return** *Word_Vector*, *WordIndex_vector*;

We process a query using the Algorithm LDPMap_Query. When a query process starts, we first build a word similarity matrix between the query term and the word vector (Line 1-5), using the *WordSimilarity* function defined above. Then we build a concept score vector between the query term and 6.87 million UMLS Metathesaurus concepts (Line 6-8). The construction of the concept score vector uses the *WordSimilarityMatrix* built previously so that there are no more word similarity calculations. In addition, it adopts a dynamic programming approach in Function *ConceptSimilarityScore*, owing to the suboptimal structure of the *ConceptSimilarity* function.

Algorithm LDPMap_Query (*query_term*)

1: **for** *i*=1: length (*query_term*)

2: **for** *j*=1: length (*Word_Vector*)

3: *WordSimilarityMatrix*[*i, j*] = *WordSimilarity*(*query_term*[*i*], *Word_Vector*[*j*]);

4: **endfor**

5: **endfor**

6: **for** *i*=1:length(*Metathesaurus*)

7: *ConceptScore_Vector*[*i*] = *ConceptSimilarityScore*(*WordIndex_vector*[*i*]);

8: **endfor**

9: **return** Concepts in Metathesaurus corresponding to top scores in *ConceptScore_Vector*;

Function ConceptSimilarityScore (*WordIndex*)

1: **for** i=2:x+1

2: **for** j=2:y+1

3: $S(i, j) = WordSimilarityMatrix[i\text{-}1, WordIndex[j\text{-}1]]$;

4: **if** $S(i, j)+S(i\text{-}1, j\text{-}1) > max\ (S(i\text{-}1, j), S(i, j\text{-}1))$;

5: $S(i, j)= S(i, j)+S(i\text{-}1, j\text{-}1)$;

6: **else if** $S(i\text{-}1, j) > S(i, j\text{-}1)$

7: $S(i, j)=S(i\text{-}1, j)$;

8: **else**

9: $S(i, j)=S(i, j\text{-}1)$;

10: **endif**

11: **endfor**

12: **endfor**

13: **return** $2*S(x+1, y+1) / (x+y)$;

A Running Example

To facilitate the understanding of our method, we provide a simple running example of our method in Tables 1 and 2. Assume the input query term is "gastro reflux". The Algorithm LDPMap_Query will first build a WordSimilarityMatrix between this query term and the word vector of Metathesaurus. Results were partially shown in Table 1.

After the WordSimilarityMatrix is available, the Algorithm LDPMap_Query will calculate the concept similarity scores between the query term and UMLS concepts by dynamic programming. The calculation will refer to WordSimilarityMatrix for word similarity score instead of calculating it again. An example of a concept similarity calculation is given in Table 2.

Complexity Analysis

The LDPMap method is much faster than the classic LCS-based word similarity calculation by treating the query term and each UMLS concept as one single word, as demonstrated in our empirical study. The classic LCS-based word similarity calculation uses dynamic programming on a character basis while we use two layers of dynamic programming, one on a character basis and the other on a word basis. To understand the analytical reason behind this speedup, let us make some simple assumptions. Assume the UMLS Metathesaurus contains M unique concepts, and each concept or query term contains t words, and each word has d characters. Also assume UMLS Metathesaurus contains K unique words. Then, the classic LCS-based word similarity calculation takes approximately $O(t^2d^2M)$ time to handle a query. However, LDPMap method takes approximately $O(td^2K+t^2M)$ time to handle this query. It is easy to observe that $K<<tM$. This explains why LDPMap is much efficient. In the following, we will see that our LDPMap approach can be further sped up with the pipeline technique.

Speeding up LDPMap with the Pipeline Technique

In building the *WordSimilarityMatrix* and *ConceptScore_Vector*, the dynamic programming method has been used for around 1.87 million times and 6.87 million times, respectively. It is interesting to find out if there are repeated calculations that can be reused to speed up the LDPMap method. By studying both the word vector and

the Metathesaurus, we found the former has a lot of repeated prefixes among words (e.g. words “4-Aminophenol”, “4-Aminophenyl”), and the latter has a lot of repeated prefix words among concepts (e.g. C1931062 ectomycorrhizal fungal sp. AR-Ny3, C1931063 ectomycorrhizal fungal sp. AR-Ny2). Thus, by lexicographically sorting the word vector and the Metathesaurus, we can use this information to save a lot of calculation in the LDPMap approach as follows:

(1) In calculating *WordSimilarityMatrix*, Given a word *A*, if it has *p* common prefix letters with the previous word *B*, the dynamic programming only needs to start from *p*+1 iteration because the previous *p*+1 columns of the dynamic programming table are exactly the same as the previous results.

(2) In calculating *ConceptSimilarityScore*, Given a concept α, if it has *q* common prefix words with the previous concept β, the dynamic programming only needs to start from *q*+1 iteration because the previous *q*+1 columns of the dynamic programming table are exactly the same as the previous results. That means, the for loop in Line 2 of Function *ConceptSimilarityScore* shall start with $j=q+2$.

The mechanism of the speedup technique can be described as a pipeline technique because a computation result can be passed down and partially reused by the subsequent computation. In the empirical study, we will see that the pipeline technique significantly improves the LDPMap speed.

A Comprehensive Query Workflow Using LDPMap Approach

Given the above solutions to the concept similarity problem under Constraints 1 and 2, we will design a comprehensive query workflow for mapping a query term to UMLS concepts. Our query workflow needs to consider multiple types of input variations and errors. Other than missing words and words in different orders that can

be properly handled by concept similarity problem formulation, we need to consider another situation when two words are merged together. In this situation, the concept similarity modelling does not fit well because it is on a word basis. Therefore it is preferable to use the classic LCS method. However, as we pointed out above, the classic LCS method is too slow for the UMLS Metathesaurus. Fortunately, we found that we can leverage concept similarity solutions, outputting a list of words with similarity score great than a threshold. When we set the threshold to be 0.35, in most cases it is able to output concepts that are similar with the query term regardless of the word merging issues. The number of outputted concepts is much smaller than the size of UMLS Metathesaurus; thus applying the LCS method on this small subset is much faster than on the whole UMLS Metathesaurus. The query workflow is illustrated in Figure 1.

In the query workflow, we first calculate concept similarity scores under Constraint 2 between the query term and all UMLS concepts. If there are concepts with scores higher than threshold T_1, we output the results and the query completes. Otherwise, we save any concepts with scores higher than threshold T_2 as $SET(T_2)$, and then perform two additional queries: (1) calculate word similarity between the query term and each concept in $SET(T_2)$ by treating the query term and each concept as one single word; (2) calculate the concept similarity scores under Constraint 1 between the query term and all UMLS concepts. Finally, we merge and output the results from (1) and (2). The number of results outputted is adjustable. An application can choose to output concepts with scores higher than a threshold, or only the top ranked concepts.

Results

To understand the actual performance of LDPMap, we implemented it in C++, and subjected it to two sets of empirical studies. In summary, the results demonstrate that LDPMap method performs much better than available methods in terms of query speed and effectiveness. All experiments were carried out on Linux cluster nodes with 2.4GHz AMD Opteron processors. For the LDPMap query workflow, we set two parameters T_1=0.8 and T_2=0.35.

Query Speed Comparison

We would like to know how fast LDPMap handles query in comparison with the standard LCS method which treats the query term and each UMLS concept as a single word, and how effective the pipeline technique for the LDPMap is. Therefore, we test the three algorithms, LCS standard, LDPMap (LDPMap_Query Algorithm) without the pipeline technique, and LDPMap algorithm with the pipeline technique, on four sets of medical concepts randomly chosen from the UMLS Metathesaurus. The first set consists of 1000 single-word medical concepts. The second, third and fourth sets consist of 1000 two-word, 1000 three-word, and 1000 four-word concepts, respectively. The results are shown in Figure 2.

From Figure 2 we can observe that the LDPMap algorithm is much faster than the standard LCS. In addition, the standard LCS method is susceptible to the word numbers in a query term while the LDPMap method is much more stable. This result is consistent with the above complexity analysis. In addition, the LDPMap with the pipeline technique significantly speeds up the basic LDPMap method. This confirms our intuition that the pipeline technique saves huge amounts of redundant computation thus improving the efficiency of the LDPMap method. As a result, we can see that in this set of experiments LDPMap with pipeline techniques on average answers a query in less than 1 second. However, the standard LCS method takes about

1 to 2 minutes in answering a query, which makes it virtually unacceptable for many biomedical applications, which can require near real-time responses, or when processing large amounts of data. In addition to the slow query time, the standard LCS is not good at processing query terms with missing words or words in different orders, as we have discussed above.

It is worthwhile to note that even for one word query, LDPMap method is significantly faster than LCS, though the concept similarity is exactly the same as the word similarity in this case. This is because the LDPMap pre-processed the UMLS terms on a word basis and built an efficient index. The similarity measurement is not directly on the UMLS terms but on words and the index which saves a lot of computational cost. In contrast, the LCS will handle the similarity measurement directly over every UMLS term. This can also be explained by our complexity analysis above. When t=1 (t is the number of words in a query), LCS complexity is $O(d^2M)$ while the LDPMap is $O(d^2K+M)$. Since $K<<M$, we conclude that LDPMap is much faster than LCS.

Next, we would like to know how effective LDPMap handles queries, especially when the query terms are slightly different than the terms in the UMLS Metathesaurus.

Query Effectiveness Comparison

To understand how effective LDPMap (referring to LDPMap query workflow in this set of experiments) handles queries with name variations and errors, we used two available methods, UMLS Metathesaurus Browser and MetaMap as benchmarks. In a cursory examination of cTAKES, we found that it exhibited similar characteristics to MetaMap in its ability to handle name variations and errors and therefore we have excluded it from comparison. Since the study on UMLS Metathesaurus Browser

requires manually inputting terms and checking the results, we have to limit the query test to manageable numbers. In addition, since the UMLS Metathesaurus Browser cannot accept a query term with more than 75 characters, we limit all query terms in our test to be no more than 75 characters. Given the above situations, and considering the fact that more than 50% of UMLS concepts contain at least 32 characters, we randomly chose 100 medical concepts with 32-75 characters from the UMLS Metathesaurus.

The 100 medical concepts are divided into two groups. The first group consists of 50 concepts with no special characters (i.e., characters other than letters and numbers), and the second group contains 50 concepts with 5 or more special characters. The two groups are for two different testing purposes.

Group 1: We will use group 1 to test how effective the query workflow handles pure English name terms, and English name terms with input errors, variations, and typos. Thus, in addition to querying the original names, we also query the names with 1, 2, 3, and 4 character variations. Character variations are generated randomly in this study, including (1) deleting a character, (2) replacing a character, (3) merging two words, i.e., deleting the white space between two words.

Group 2: We will use group 2 to test how effective the query algorithm is in handling many professional medical terms, which may contain a good number of special characters, such as chemical compounds and drugs. To simulate the name variations that frequently appear in these terms, we randomly apply 1, 2, 3, and 4 character variations, including (1) deleting a special character, (2) replacing a special character by a white space.

To complement the above test groups, we use the following group to test how effective the query algorithm handles short terms which may be queried commonly in real situation.

Group 3: We randomly picked 100 medical concepts with 5-31 characters. Since many of these concepts are quite short, we only apply 1 and 2 random character variations, including (1) deleting a character, (2) replacing a character, (3) merging two words.

In these experiments, we found that MetaMap often output multiple matching results but there are no ranks of these results. In contrast, the UMLS Metathesaurus Browser usually outputs a list of ranked concepts, and LDPMap can be configured to output the top k (k>=1) ranked concepts.

Thus, to be as fair as possible, we use two criteria to measure the correctness of a query:

Criterion 1: A query is correct if the original term appear (1) in top 25 ranked concepts (i.e., in the first page of the result) by the UMLS Metathesaurus Browser; (2) in the top 25 ranked concepts by LDPMap; (3) in the result of MetaMap.

Criterion 2: A query is correct if the original term appears (1) as the top ranked concept by UMLS Metathesaurus Browser; (2) as the top ranked concept by LDPMap.

Criterion 1 indicates if the query processing mechanism is able to handle the query with reasonable accuracy. Criterion 2 is much stringent and it indicates whether a method can be applied to applications require high accuracy.

Figures 3 and 4 are the error rate for the two groups of experiments, under Criterion 1. From both figures, we can clearly see that the LDPMap approach has very few errors among all tests. In comparison, the UMLS Metathesaurus Browser and MetaMap's

error rate are quite high especially when multiple characters changes are present. MetaMap has a considerable error rate even when querying the original terms (0 characters changes). This may owe to the text processing mechanism of MetaMap. Since MetaMap is targeted at finding medical terms from a biomedical text, it leverages a combination of part-of-speech tagging, shallow parsing, and longest spanning match against terms from the SPECIALIST Lexicon before matching terms against concepts in the UMLS. Therefore, it tends to decompose longer spans of text and medical terms into several shorter medical terms.

Figures 5 and 6 are the error rates for the two groups of experiments, under Criterion 2. Since MetaMap usually outputs multiple concepts without ranking, we exclude MetaMap from the Criterion 2 measurement. From these two figures, we can observe that the error rate of the UMLS Metathesaurus Browser is much higher in comparison with the measurement of Criterion 1. Quite surprisingly, there are some errors even when querying a few original terms (such as " Distal radioulnar joint"). This suggests that UMLS Metathesaurus Browser is not suitable for query processing for applications that have a high-accuracy demand. In contrast, the LDPMap still has a very low error rate, on average less than 5% across the 0-5 character changes, and free of errors in querying the original terms.

From Figures 7 and 8, we can see that the general performances of LDPMap, UMLS Metathesaurus Browser, and MetaMap on short query terms are similar to their performances on long query terms. LDPMap still has a clear advantage over UMLS Metathesaurus Browser, and MetaMap. However, we noticed that LDPMap error rate reaches 27% for 2 character changes under Criterion 2. This is understandable because generally short terms contain fewer words than long terms, and the concept similarity measurement is less favoured. However, the parameter T_1 can be used as an

adjustment of preference between the concept similarity measurement and the word similarity measurement. By increasing T_1 from 0.8 to 0.85, we observed that this error rate reduces from 27% to 20%. This demonstrates that LDPMap is flexible in handling both long and short term queries.

To provide some details on the medical concepts we used in this set of experiments, and the character changes applied. We list a few of them in Tables 3. From this table, we can see that it contains concepts of different lengths. The randomly generated character variations cover several common cases of text data inaccuracy, including, misspellings, merging of two words, and special character omissions. From Table 4 we can see that MetaMap cannot handle them properly. Instead, it finds some concepts related to individual words in the query term. The UMLS Metathesaurus Browser does not do any better on them. In contrast, LDPMap correctly answered all these queries except for "AlbunexIectable Product". Although "Injectable Product" is not correct, it is at least closer to the original term than those returned by the UMLS Metathesaurus Browser and MetaMap. By reviewing the LDPMap approach, we conclude that this error can be eliminated if we increase the threshold T_1 to a value such that word similarity (LCS) is used to measure the two terms. To confirm this, we increase T_1 from 0.8 to 0.85, and LDPMap successfully returns the original term. However, a high T_1 implies that LDPMap gives more preference to LCS-based similarity measurement than to concept similarity measurement defined above. Consequently, LDPMap will be less productive in handling real-world queries that contain incomplete medical terms (i.e., medical terms with missing words). It is quite evident that there does not exist one set of T_1 and T_2 that fits all situations. As a result, we will fine tune these parameters to leverage LDPMap in our future applications.

Conclusions

In the work we proposed LDPMap, a layered dynamic programming approach to efficiently mapping inaccurate medical terms to UMLS concepts. As a main advantage of the LDPMap algorithm, it runs much faster than classical LCS method therefore makes it possible to efficiently handle UMLS term queries. When similarity is counted on a word basis, LDPMap algorithm may yield a more desirable result than LCS. In other cases (such as word merging), it is possible that LCS query results are more preferable. Thus, in the comprehensive query workflow of LDPMap, the LDPMap method is complemented by LCS and adjustable by parameter T_1. Different from using LCS alone, the LDPMap query workflow only applies LCS (when needed) to a very limited number of candidate terms thus achieves a very fast query speed.

In query effectiveness comparison, we observed that LDPMap has a very high accuracy in processing queries over the UMLS Metathesaurus involving inaccurate terms. In contrast, the UMLS Metathesaurus Browser has a very limited ability in handling these queries, though it can handle queries of accurate terms fairly well. Throughout the study, we also observed that MetaMap, in general, is not suitable for mapping long medical terms to the UMLS concepts as it focuses on extracting short medical terms from the query text.

Although LDPMap is very efficiently in handling UMLS term queries, it has two major limitations. First, it cannot handle synonyms and coreferences. Fortunately, UMLS Metathesaurus often list a concept preferred names and synonyms so that LDPMap can work effectively in most cases, though the list may still not be complete. Second, it is not able to perform syntax-level processing as MetaMap does, such as extracting medical terms from an article. Whether it is possible to extend the LDPMap approach to overcome the two limitations remains an open question. In the future we would like to investigate this question and plan to use LDPMap as an efficient pre-

processing tool to map medical terms to the UMLS concepts, and use the results in our knowledge discovery platform.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

KR implemented the LDPMap algorithm, carried out the experiments, and edited the manuscript. AL, AM, RM, and KH analyzed comparable methods, participated in the design of the study, and revised the manuscripts. YX led the project including development of the idea, design of the algorithms, and writing of the manuscript.

Acknowledgements

AL and AM were supported by award number R01LM011116 from the National Library of Medicine. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Library of Medicine or the National Institutes of Health.

KH was supported in part by the Department of Defense CDMRP Grant (CA100865).

References

1. Aronson, A.: **Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.** In : *AMIA Symposium*, p.17 (2001)

2. Savova, G., Masanz, J., Ogren, P., Zheng, J., Sohn, S., Kipper-Schuler, K., Chute, C.: **Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications.** *Journal of the American Medical Informatics Association* 17(5), 507-513 (2010)

3. Bodenreider, O.: **The unified medical language system (UMLS): integrating biomedical terminology.** *Nucleic Acids Res. 32(Database issue)*, D267-D270 (January 2004)

4. Xiang, Y., Lu, K., James, S., Borlawsky, T., Huang, K., Payne, P.: **k-neighborhood Decentralization: A Comprehensive Solution to Index the UMLS for Large Scale Knowledge Discovery.** *Journal of Biomedical Informatics* 45(2), 323-336 (2012)

5. Payne, P., Borlawsky, B., Lele, O., James, S., Greaves, A.: **The TOKEn project: knowledge synthesis for in silico science.** *Journal of the American Medical Informatics Association* 18(Suppl 1), i125 (2011)

6. Melton, G. B., Parsons, S., Morrison, F. P., Rothschild, A. S., Markatou, M., Hripcsak, G.: **Inter-patient distance metrics using SNOMED CT defining relationships.** *Journal of Biomedical Informatics* 39(6), 697--705 (2006)

7. McInnes, B. T., Pedersen, T., Pakhomov, S. V. S.: **UMLS-Interface and UMLS-Similarity: Open source software for measuring paths and semantic similarity.** In : *AMIA Annual Symposium Proceedings*, p.431 (2009)

8. Ghali, W., Hall, R., Rosen, A., Ash, A., Moskowitz, M.: **Searching for an improved clinical comorbidity index for use with ICD-9-CM administrative data.** *Journal of Clinical Epidemiology* 49(3), 273-278 (1996)

9. Denny, J.: **Mining Electronic Health Records in the Genomics Era.** *PLoS computational biology* 8(12), e1002823 (2012)

10. Levenstein, V.: **Binary codes capable of correcting spurious insertions and deletions of ones.** *Problems of Information Transmission* 1(1), 8--17 (1965)

11. Sankoff, D., Kruskal, J.: **Time warps, string edits, and macromolecules: the theory and practice of sequence comparison.** (1983)

12. Needleman, S., Wunsch, C.: **A general method applicable to the search for similarities in the amino acid sequence of two proteins.** *Journal of molecular biology* 48(3), 443--453 (1970)

13. Apostolico, A., Guerra, C.: **The longest common subsequence problem revisited.** *Algorithmica* 2(1-4), 315--336 (1987)

14. Navarro, G.: **A guided tour to approximate string matching.** *ACM computing surveys (CSUR)* 33(1), 31--88 (2001)

15. Cucerzan, S., Brill, E.: **Spelling correction as an iterative process that exploits the collective knowledge of web users.** In : *Proceedings of EMNLP*, vol. 4, pp.293--300 (2004)

16. RU, L., WANG, C., WU, Y., MA, S.: **Search Query Correction based on User Intent Analysis.** *Journal of Computational Information Systems* 9(6), 2157--2166 (2013)

17. West, D.: **Introduction to graph theory 2.** (2001)

18. **cTAKES 2.5 User Install Instructions.** In: wiki.nci.nih.gov. Available at: https://wiki.nci.nih.gov/display/VKC/cTAKES+2.5+User+Install+Instructions#cTAKES25UserInstallInstructions-BundledUMLSDictionaries

19. Jimeno, A., Jimenez-Ruiz, E., Lee, V., Gaudan, S., Berlanga, R., Rebholz-Schuhmann, D.: **Assessment of disease named entity recognition on a corpus of annotated sentences.** *BMC bioinformatics* 9(Suppl 3), S3 (2008)

Figures

Figure 2

Query time of LCS, LDPMap and LDPMap pipeline on randomly chosen 1000 medical concepts.

Figure 3

Correctness comparison on LDPMap, UMLS Metathesaurus Browser, and MetaMap for Group 1 using Criterion 1.

Figure 4

Correctness comparison on LDPMap, UMLS Metathesaurus Browser, and MetaMap for Group 2 using Criterion 1.

Figure 5

Correctness comparison on LDPMap and UMLS Metathesaurus Browser for Group 1 using Criterion 2.

Figure 6

Correctness comparison on LDPMap and UMLS Metathesaurus Browser for Group 2 using Criterion 2.

Figure 7

Correctness comparison on LDPMap, UMLS Metathesaurus Browser, and MetaMap for Group 3 using Criterion 1.

Figure 8

Correctness comparison on LDPMap and UMLS Metathesaurus Browser for Group 3 using Criterion 2.

Tables

Table 1.

An example of *WordSimilarityMatrix* constructed for query term "gastro reflux".

		***Word_Vector* of Metathesaurus**						
		…	gastro (at i)	…	Oesophageal (at j)	…	reflux (at k)	…
Query term	gastro	…	1 (gastro)	…	0.235294 (so/ga)	…	0.166667(r)	…
	reflux	…	0.166667 (r)	…	0.235294 (el)	…	1(reflux)	…

Table 2

An example of calculating the concept similarity score between the query term "gastro reflux" and the UMLS concept "gastro oesophageal reflux" for the *ConceptScore_Vector* construction. The calculation will refer to the *WordSimilarityMatrix* as shown in Table 1. The normalized final similarity score is 2*2/(2+3)=0.8.

		UMLS concept	gastro	oesophageal	reflux
		word index	i	k	j
query term	order		0	0	0
gastro	1	0	1	1	1
reflux	2	0	1	1.23594	2

Table 3

Original terms and their randomly generated character variations

CUI	name	Randomly generated 4 character variations
C3267394	POMEGRANATE FRUIT EXTRACT 150 MG Oral Capsule	POMGRAATE FRUIT EXTRdCT 150 MG Oral Casule
C3228202	Albunex Injectable Product	AlbunexIectable Product
C0505183	Lateral branch of dorsal ramus of fifth thoracic spinal nerve	LateMa branch of dorsal ramus of ifth thoracic gpinal nerve
C1459293	Sinorhizobium americanus	Sinokhizrbimamericanus
C1541607	gp100/IL-7/ISA-51/MART-1	gp100 IL 7ISA-51/MART1
C1352046	danthron 1.5 MG/ML / Pantothenic Acid 2.5 MG/ML Oral Suspension	danthron 15 MGML Pantothenic Acid 25 MG/ML Oral Suspension
C0040372	Benzenesulfonamide, N-(((hexahydro-1H-azepin-1-yl)amino)carbonyl)-4-methyl-	Benzenesulfonamide, N-((hexahydro1H-azepin-1-yl amino)carbonyl-4-methyl-
C2714409	1-undecene-1-O-beta-2',3',4',6'-tetraacetyl glucopyranoside	1-undecene1-O-beta2,3',4',6-tetraacetyl glucopyranoside

Table 4

Query results for Table 3.

CUI	UMLS Metathesaurus Browser (concept ranked 1st by approximate match)	MetaMap	LDPMap
C3267394	C0030054 Oxygen	C0016767 Fruit, C2346927 Mg++, and 4 others	correct
C3228202	C1514468 product	C1704444 Product (Multiplicative Product) [Quantitative Concept] C1514468 product [Entity]	C0086466 Injectable Product
C0505183	C0007965 Chediak-Higashi Syndrome	C1706131 Branch(Branch(group)), C2700383 Branch(Branch of plant), and 6 others	correct
C1459293	No result	No result	correct
C1541607	C1512807 Integrated Learning System	C0020898 IL (Illinois (geographic location)), C1522481 MART-1 (MART-1 Tumor Antigen) , and 2 others	correct
C1352046	C0029383 Osmium	C1129294 danthron 25 MG, C0439526 /mL [Quantitative Concept], and 3 others	correct
C0040372	C0265215 Meckel-Gruber syndrome	C0053169 benzenesulfonamide, C0441922 N+ (N+ (tumor staging)), and two others	correct
C2714409	C0030011 Oxidation	C0470206 +1 [Q uantitative Concept] C1417683 BETA2 (NEUROD1 gene), and 7 others	correct